DOCUMENT RESUME

ED 069 684 TM 002 135

AUTHOR Ruch, William W.
TITLE Statistical, Legal, and Moral Problems in Following the EEOC Guidelines.
PUB DATE 21 Apr 72
NOTE 19p.; Paper presented at the annual meeting of Western Psychological Association (Portland, Oregon, April 21, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Aptitude Tests; *Civil Rights; Decision Making; Employment Problems; Job Skills; Legal Problems; Minority Groups; Moral Issues; Predictive Ability (Testing); *Predictive Measurement; Predictor Variables; *Racial Differences; Speeches; Statistical Analysis; *Test Bias; Testing; Test Interpretation; Test Validity; *Vocational Aptitude

ABSTRACT

Statistical, legal and moral problems involved in following the EEOC guidelines are described. The guidelines require separate data for minority and non-minority groups with differential cut off scores for aptitude tests which have a racial bias. Problems reviewed include: identification of racial bias in tests is difficult; giving one race an advantageous cutoff over another may be unfair, creating legal challenges; and determining selection by race may diminish the effectiveness of the work group. The author suggests selection on the basis of proportion of numbers of each race applying, taking the top from each group. (DJ)

STATISTICAL, LEGAL, AND MORAL PROBLEMS IN FOLLOWING THE EEOC GUIDELINES

by

William W. Ruch

Psychological Services, Inc.

Los Angeles, California

Presented at

the Symposium on:

"Differential Validity and EEOC Testing Guidelines"

Western Psychological Association

Portland, Oregon

April 28, 1972

Statistical, Legal, and Moral Problems In Following the EEOC Guidelines[1]

William W. Ruch
Psychological Services, Inc.
Los Angeles, California

The Guidelines (4) state that, "Data must be generated and results separately reported for minority and nonminority groups whenever technically feasible....A test which is differentially valid may be used in groups for which it is valid but not for those in which it is not valid. In this regard, where a test is valid for two groups but one group characteristically obtains higher test scores than the other without a corresponding difference in job performance, cutoff scores must be set so as to predict the same probability of job success in both groups."

This requirement is apparently based upon the assumption that there is likely to be a difference among racial or sex subgroups in the applicant population with respect to the regression line by which a criterion of job performance is predicted from test scores. There is increasing reluctance on the part of knowledgeable psychologists to make this assumption, at least, with regard to black-white comparisons. For example, Bray and Moses state on page 554 of

---

[1] Presentation in symposium on "Differential Validity and EEOC Testing Guidelines" at the annual meeting of the Western Psychological Association, Portland, Oregon, April 21, 1972.

their authoritative article in the current Annual Review of Psychology:

> "Do aptitude test scores, obtained under proper conditions of administration, show significantly different validities for minority and majority group members in predicting a pertinent measure of job proficiency? This question is still open since there are few such studies. It does appear, however, that the closer the study design comes to the ideal, the less likelihood there is of finding differential validity."

However, this is another topic, so I won't dwell on it.

Unfortunately, the Guidelines do not give guidance with respect to what inferential techniques should be used in determining when a single regression line does not apply to all groups. There is an implication in the Guidelines that "differentially valid" means valid for one group but not for another. The important situation in which a test is valid, but not equally valid, for two groups is not covered. When this requirement of separate validation for minority and nonminority groups is taken in conjunction with the requirement stated two paragraphs later that the obtained correlation coefficient be statistically significant at the 5% level, an unwary researcher might be lead down the primrose path of calling a test differentially valid if for one group the null hypothesis of r = zero is rejected at the 5% level and for another group the null hypothesis is not rejected. While such

an approach may seem reasonable at first blush, a careful consideration of its implications will show that it is unworkable. For any population in which the correlation between two variables is other than zero, the finding or non-finding of statistical significance in a sample is a function of both the size of the correlation in the population and of the number of cases in the sample. As either of these increases the probability of rejecting the null hypothesis at any stated significance level increases. If a large sample and a small sample are taken from a single population in which the correlation coefficient is greater than zero, the probability of obtaining statistical significance is greater in the large sample than in the small sample. If we follow the significant-for-this-group-but-not-for-that-group strategy, we are stacking the deck in favor of a finding of differential validity since in the typical, real-life situation the sample size for whites will be considerably greater than the sample size for blacks. Several examples of this are to be found on page 132 of Testing and Fair Employment, by Kirkpatrick et al (5), which is reproduced in Exhibit I of the handout. Applying the 5% level of significance, there are eight instances in which the obtained validity for whites is statistically significant but the obtained validity for Negroes is not. Using the 1% level, there are nine such instances. In five comparisons, the validity for whites is significant at the 1% level, but the validity for Negroes fails to reach significance

even at the 5% level. Yet, there is no basis for inferring that the validity of the tests is anything but the same for the two groups. Take a look at the proverbs test as a predictor of the salary criterion. For whites, the validity is .13, significant at the 1% level; for Negroes the validity is .16, which although higher than that for whites, is not even significant at the 5% level. Obviously, the significant-for-this-group-but-not-for-that-group strategy fails to yield credible inferences at least in the present instance. As a matter of fact, Kirkpatrick et al (5) conclude from this table that "Perhaps the most important finding of the present study is the similarity of validity coefficients for both ethnic groups."

The Guidelines state that "a test which is differentially valid may be used in groups for which it is valid but not for those in which it is not valid." It would clearly be incorrect to adopt a policy of using these tests for whites, but not for Negroes, solely because of the findings presented in Exhibit I.

The other consideration in comparing the regression lines of two or more subgroups is that of fairness. Even if the test predicts equally well for two groups it may, on the average, underestimate the job performance of one group and overestimate the job performance of another. The only guidance the Guidelines give us in this respect is that "where a test is valid for two groups but one group characteristically obtains higher test

scores than the other, without a corresponding difference in job performance, cutoff scores must be set so as to predict the same probability of job success in both groups." We are left without operational procedures for determining when between-group differences in criterion scores correspond to between-group differences in test scores. Additionally, the Guidelines provide us with no justification whatsoever for applying different cutoff scores for the two groups in the event that the average test scores are the same but there is a difference in average criterion scores.

Here again one might be tempted to compare the results of one significance test with the results of another. A stated significance level - say 5% - could be established and t-tests could be run between criterion means and also between test means. If a significant difference were found between the test means of whites and blacks, but a significant difference were not found between their criterion means, it would be concluded that the tests were unfair to the group with the lower test scores and that there was sufficient statistical evidence to warrant the use of different cutoff scores. Yet this conclusion could easily be in error. Suppose the difference in test means were significant at the .05 level and the difference in criterion means were significant at the .06 level. An employer who based his decision to use differential cutoff scores on such flimsy evidence would be inviting a successful lawsuit from a member of the group for whom the higher cutoff score was required.

What is needed is a single significance test of the null hypothesis that the regression line in the population of whites is colinear with the regression line in the population of blacks. Such a test is accomplished by using analysis of covariance. A regression line can be defined in terms of its slope and its y-intercept. If two or more regression lines have the same slope and have the same intercept, they must be colinear. If they have different slopes, then there is a difference in validity between the groups. If the regression lines have different intercepts, then there is a lack of correspondence of test means and criterion means between groups. Differences in slopes indicate differential validity; differences in intercepts indicates unfairness. Depending on the analysis of covariance model used, the significance of the difference between slopes and between intercepts can be assessed either separately or together. Predictors can be studied one at a time or combined in a multiple regression equation. If an analysis of covariance results in significant differences in slopes and/or intercepts, the same cutoff score should not be used for both groups.

Aside from the statistical problems involved in making correct inferences with respect to the regression lines of two or more subpopulations, there are important moral, and legal problems to be wrestled with. Before getting into them, let's take a look at regression lines under different conditions of equality or inequality of slopes and intercepts. Two straight lines in a plane can have just three relationships between them: They can be colinear; they can be parallel; or they can intersect. These three situations are depicted in Cases 1, 2, and 3 of Exhibit II of the handout. When predictor means and standard deviations and criterion means and standard deviations are each free to vary independently of the others, there are several possible configurations within each case. For purposes of illustration in the remainder of this

presentation I have assumed that the predictor standard deviations and the criterion standard deviations are equal for the two groups. For Case 3, I will talk just about the situation in which both slopes are positive, but bear in mind that this will include the situation in which the smaller slope is so small as to be essentially zero. For each case I will consider just two subproblems, one in which the criterion means are the same for the two groups, the other in which the criterion means differ. Note that in Case 1, in which there is a single regression line, when the criterion means for the two groups are equal then the test means are of necessity equal; when the criterion means are unequal then the test means must be unequal. In Case 2, parallel regression lines - equal slopes, unequal intercepts - when the criterion means are equal, the test means must be unequal. There are several other subproblems, particularly in Case 3, but the two which are given will suffice for the purposes of my illustration.

In separate articles in the Summer, 1971, issue of the Journal of Educational Measurement, both Thorndike (6) and Darlington (3) pointed out that a policy of using tests in such a manner as to maximize fairness will sometimes conflict with the policy of using them to maximize validity. This can be seen from the figure in Exhibit II. First, let's define terms. Cleary (2) has given the definition:

> "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup."

This leads to a finding of bias whenever the regression lines are parallel and to bias as a function of test score, whenever the lines intersect. However, rather than talking about the bias of a test my purpose here is to talk about the fairness of the use of a test. In the employment situation selection decisions are ultimately dichotomous - either the applicant is hired or he is not.

One definition of the fair use of a test which has been advanced is that a test is used fairly if decisions are made on the basis of the predicted criterion score, and when separate prediction equations are used when appropriate. Let's apply this definition to Case 3-A. Suppose we select only those applicants for whom the criterion is predicted to be at least fifty-four. The regression equation for whites would be $Y' = 50 + .40(X-50)$. To have a predicted criterion score of fifty-four, a white applicant would need a test score of sixty. Sixteen percent of white applicants would meet this standard. The regression equation for blacks is $Y' = 50 + .20(X-50)$. To have a predicted criterion score of fifty-four, a black applicant would need a test score of seventy. Two percent of black applicants would meet this standard. Thus, under Case 3-A in which the tests are valid for both whites and blacks, but are more valid for whites, and in which blacks and whites perform equally well on the job, selecting on the basis of the predicted criterion score results in the

selection ratio for whites being eight times the selection ratio for blacks. Black applicants would be penalized, so to speak, by virtue of belonging to a less predictable group. Try to get that one by a federal judge, much less Bill Enneis. I will be quite interested in what Bill has to say about this. The Guidelines are silent as to what to do in such a situation. In a different situation, depicted here as 2-A, they state that "cutoff scores must be set so as to predict the same probability of job success in both groups." This is essentially the same as selecting on the basis of predicted criterion scores. Although this strategy is appropriate for Case 2-A, it results in what most of us would call unfairness in Case 3-A. A definition of the fair use of a test which is more reasonable to me than is the practice of hiring on the basis of predicted criterion score or on the basis of predicted probability of job success is one which has been set forth by Thorndike (6). One of his definitions of fair use of a test is "providing each group the same opportunity for admission to training or to a job as would be represented by the population of the group falling above a specified criterion score on the correlated variable of training or job performance." In other words, if we hired every applicant and then defined job success in terms of reaching or surpassing some specified criterion score we could then determine what

percentage of successful job performers were black, what percent were white and so on. Under this definition of fair use of a test, if we found that 17% of the successful job performers were black, then we should adjust our cutoff scores so that 17% of those selected are black. If, in an unselected group, 5% of the successful job performers are black, then our cutoff scores should be so arranged that 5% of the people passing are black. Another way of stating this same definition is that the percentage of blacks among the selected group should be equal to the percentage of blacks among the group which would be selected on the basis of a test of perfect validity. I will use this definition for the rest of my presentation.

Next, we need a definition for maximum validity. The one that I will use is simply that for a given selection ratio validity is maximized when the mean criterion score of selectees is maximized. In other words, the selection strategy with the highest validity is the one that selects people with the best job performance.

Now let's consider the results when we apply different selection strategies to these different models. The first strategy will be what, until recent years was the most common one in industry. That is the use of the same cutoff score for all applicants. In other words, the cutoff score is the same for all groups and the selection ratio is free to vary from group to group as a function of their test scores.

The second strategy will be one which has come into vogue in recent years and that is the applicant-based quota. Here the selection ratio is kept the same for all subgroups and the cutoff score is allowed to vary from group to group. If 20% of all applicants are to be hired, the top 20% of the whites, the top 20% of the blacks, etc. are hired.

This results in selection being apportioned among the subgroups in accordance with each subgroups' representation in the applicant population. If 11% of the applicants are black, then 11% of those selected will be black.

The third strategy will be that of separate regression equations in which each applicant is selected on the basis of his predicted criterion score, using the appropriate regression equation. Of course, when the regression lines are colinear, this would result in using the same cutoff score for all groups.

The fourth strategy I will call the success-based quota. Here, quotas are established so that the proportions of subgroups among selectees are equal to the proportion of subgroups among those who would be successful on the job if all applicants were hired. This is equivalent to our definition of the fair use of a test.

In Exhibit III of the handout, these four selection strategies are applied to the six regression situations which we discussed earlier. In Case 1-A, a single regression line with no between-

group differences in criterion means, it makes no difference which model is used. The same individuals will be selected in any event. Thus, all strategies have maximum validity. Under our definition of fairness, all strategies are fair. The proportion of blacks among selectees is equal to the proportion of blacks among successful job performers. As we would expect, in Cases 2 and 3 - in which the tests work differently for different subgroups - the use of the same cutoff score for everyone is inappropriate from the standpoint of both validity and fairness. This, of course, is what the Motorola case, the Guidelines and the entire testing controversy is all about. But let's consider some of the problems with Case 1-B. Here, a single regression line depicts the relationship between predictor and criterion for both groups, but one group has lower test scores and lower criterion scores. I would assume that this would not trigger the Guidelines section on unfairness since there is a between-group difference in test scores. However, note that there is a twelve and a half point difference in test scores but only a five point difference in criterion scores. In these illustrations, all standard deviations are equal to ten. Thus, as a necessity, if both means are to fall on the same regression line, there is far more overlap in terms of job performance than in terms of test score. Thirty-one percent of blacks are above the white mean criterion score, but only eleven percent are above the white mean test score. To work out what would happen if we applied the

same cutoff score of 50 to both groups, let's assume that 20% of the total group is black and 80% is white. Let's define successful job performance as having a score of 50 or more on the criterion. As it works out, 87% of the successful job performers are white and 13% of the successful job performers are black. Thus, under our definition of fairness, 13% of those selected should be black. However, only 5% of selectees will be black. That is, of those passing the cut-off score of 50, 5% are black and 95% are white. Under the Supreme Court rule that a test with an adverse impact must be job related, I would assume that the use of the same cutoff score or at least separate regression equations would be legal in all of these cases. As I understand the Guidelines, the use of the same cutoff score would be legal in Case 1-B. However, we do have a moral issue in Case 1-B. Is the fact that the use of a single cutoff score maximizes validity sufficient justification to have only 5% blacks on the job when 13% of those who would perform the job successfully are black? Stated another way, a perfectly valid test would yield 13% blacks among those selected, yet the test depicted in Case 1-B would yield only 5%. In order to raise this to 13% we would have to adjust the cutoff scores in accordance with the success-based quota. Yet this would lower the validity of our selection procedure and thus the efficiency of our work force.

In examining the rest of the table in Exhibit III, we find that separate regression equations will always yield maximum validity and that the success-based quota will always yield fairness. However in many situations we must make a choice between these two important goals.

I realize that I have covered some rather technical material in a very short time, and that an oral presentation such as this is difficult to follow. I hope, though, that I have convinced most of you that there are serious statistical, legal and moral problems which are not resolved by the EEOC Guidelines.

REFERENCES

1. Bray, D. W. and Moses, J. L. Personnel Selection. Annu. Rev. Psychol., 1972, 23, 545-576.

2. Cleary, T. Test bias: Prediction of grades of Negro and White students in integrated colleges. J. educ. Measmt, 1968, 2, 115-124.

3. Darlington, R. Another look at "cultural fairness." J. educ. Measmt, 1971, 8, 71-82.

4. Equal Employment Opportunity Commission. Part 1607 - Guidelines on employee selection procedures. Code of Federal Regulations, Title 29, Chapter XIV, 1970.

5. Kirkpatrick, J. et al. Testing and fair employment. New York: New York University, 1968.

6. Thorndike, R. Concepts of culture-fairness. J. educ. Measmt, 1971, 8, 63-70.

Exhibit I***

Concurrent Validity Coefficients

Test	Group	Salary Criterion	Performance Rating Criterion
Checking (1)	Total	12**	15**
	White	12*	16**
	Negro	12	16
Checking (2)	Total	24**	16**
	White	24**	17**
	Negro	23*	14
Sorting	Total	10*	08
	White	12*	12*
	Negro	05	02
Proverbs	Total	14**	04
	White	13**	04
	Negro	16	06
Vocabulary	Total	28**	17**
	White	29**	20**
	Negro	30**	13
Spelling	Total	26**	19**
	White	26**	18**
	Negro	29**	25*
Arithmetic	Total	22**	17**
	White	23**	20**
	Negro	20*	13
General	Total	23**	18**
	White	24**	17**
	Negro	25*	30**

NOTE: N for total group equals 535; N for whites equals 437; N for Negroes equals 98.
Decimal points omitted.
*p $<$ .05.
**p $<$ .01.

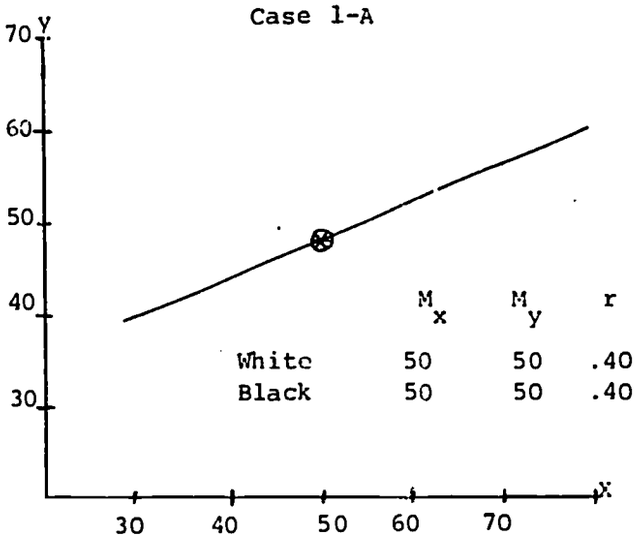
***Kirkpatrick, J. J. et al *Testing and fair employment*. New York: New York University, 1968, page 132.

EXHIBIT II

O White Mean

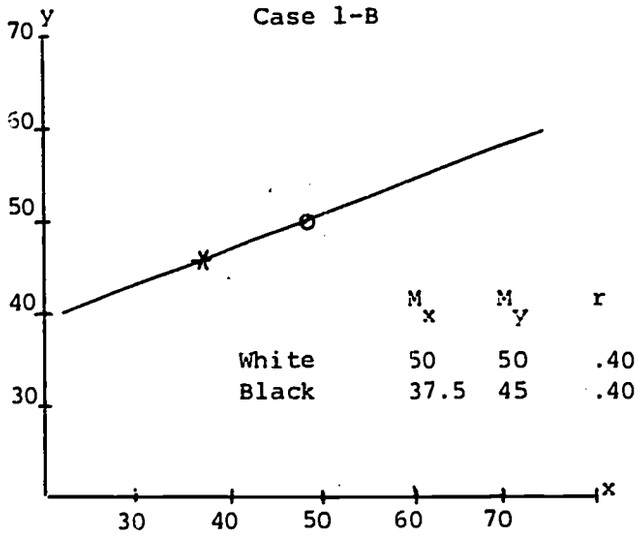
* Black Mean

Case 1-A



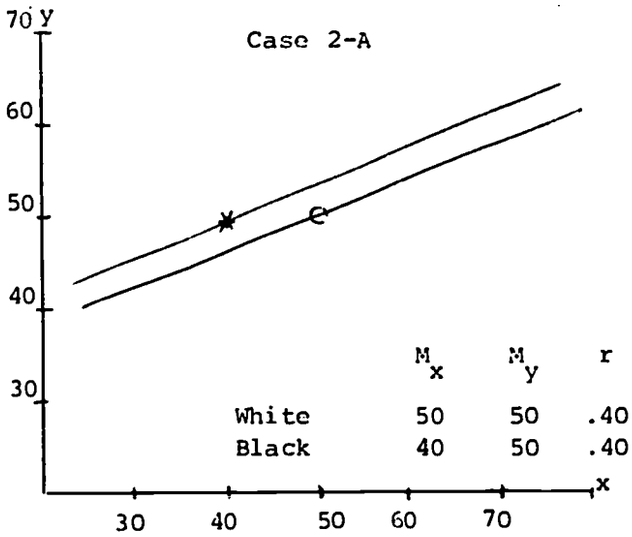
	M_x	M_y	r
White	50	50	.40
Black	50	50	.40

Case 1-B



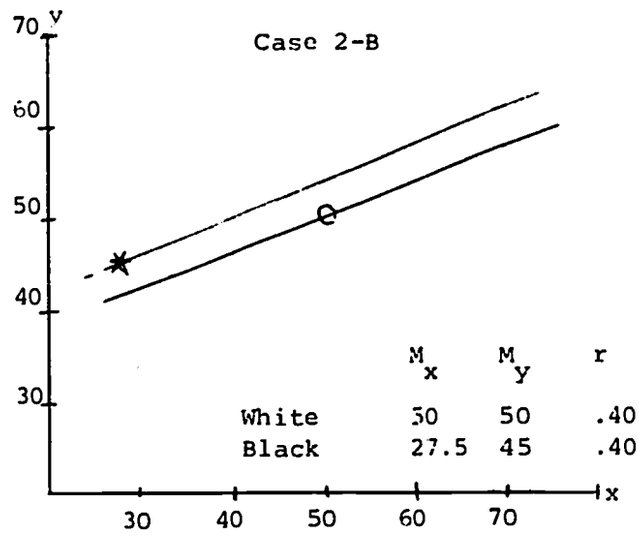
	M_x	M_y	r
White	50	50	.40
Black	37.5	45	.40

Case 2-A



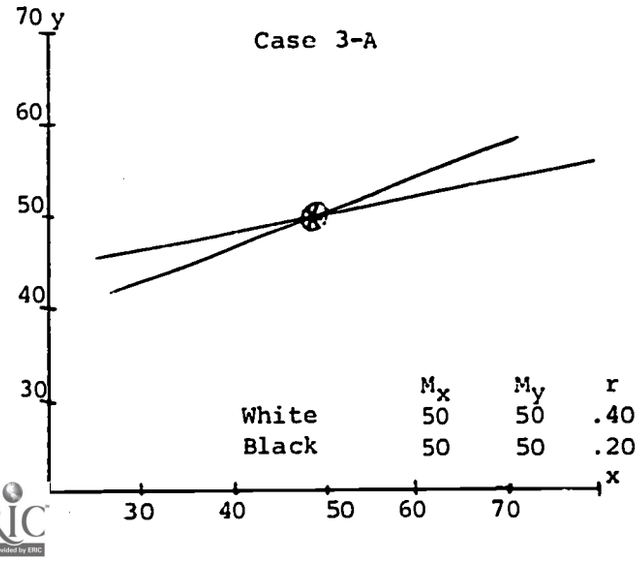
	M_x	M_y	r
White	50	50	.40
Black	40	50	.40

Case 2-B



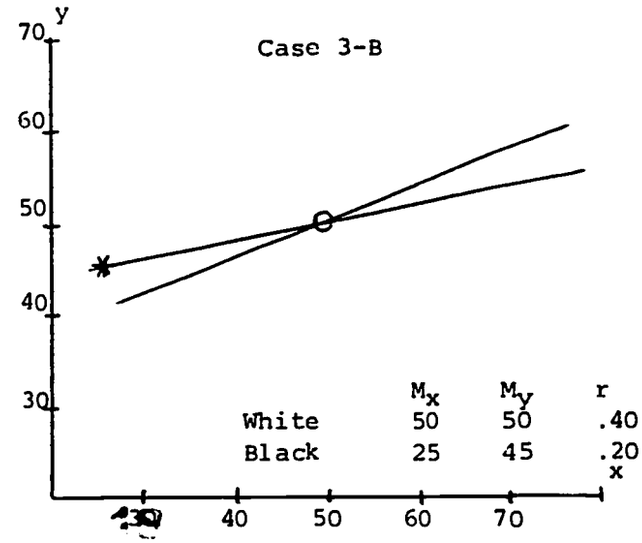
	M_x	M_y	r
White	50	50	.40
Black	27.5	45	.40

Case 3-A



	M_x	M_y	r
White	50	50	.40
Black	50	50	.20

Case 3-B



	M_x	M_y	r
White	50	50	.40
Black	25	45	.20

EXHIBIT III

APPLICATION OF FOUR SELECTION STRATEGIES TO SIX REGRESSION SITUATIONS

Case	Description	Comparison	Same Cut-Off Score	Applicant-Based Quota	Separate Regression Equation	Success-Based Quota
1-A	Single Regression Line, Equal Criterion Means, Equal Test Means	Maximum Validity	Yes	Yes	Yes	Yes
		Always Fair	Yes	Yes	Yes	Yes
1-B	Single Regression Line, Unequal Criterion Means, Unequal Test Means	Maximum Validity	Yes	No	Yes	No
		Always Fair	No	No	No	Yes
2-A	Parallel Regression Lines, Equal Criterion Means, Unequal Test Means	Maximum Validity	No	Yes	Yes	Yes
		Always Fair	No	Yes	Yes	Yes
2-B	Parallel Regression Lines, Unequal Criterion Means	Maximum Validity	No	No	Yes	No
		Always Fair	No	No	No	Yes
3-A	Intersecting Regression Lines, Equal Criterion Means	Maximum Validity	No	No	Yes	No
		Always Fair	No	Yes	No	Yes
3-B	Intersecting Regression Lines, Unequal Criterion Means	Maximum Validity	No	No	Yes	No
		Always Fair	No	No	No	Yes